Analysis of Factors Affecting System Performance in the ASpIRE Challenge

Jennifer Melot, Nicolas Malyska, Jessica Ray, and Wade Shen

MIT Lincoln Laboratory

{jennifer.melot, nmalyska, jessica.ray, swade}@ll.mit.edu

Abstract

This paper presents an analysis of factors affecting system performance in the ASpIRE (Automatic Speech recognition In Reverberant Environments) challenge. In particular, overall word error rate (WER) of the solver systems is analyzed as a function of room, distance between talker and microphone, and microphone type. We also analyze speech activity detection performance of the solver systems and investigate its relationship to WER. The primary goal of the paper is to provide insight into the factors affecting system performance in the ASpIRE evaluation set across many systems given annotations and metadata that are not available to the solvers. This analysis will inform the design of future challenges and provide insight into the efficacy of current solutions addressing noisy reverberant speech in mismatched conditions.

Index Terms: speech recognition, reverberant rooms, microphone audio

1. Introduction

The development of automatic speech recognition (ASR) that is able to perform well across a variety of acoustic environments and recording scenarios is the focus of many research efforts [1, 2]. Previous work with the AMI meetings room corpus [3], the ICSI meeting corpus [4, 5], and the MC-WSJ-AV corpus [6], for example, have shown that ASR performance degrades in various room and microphone conditions and also when data used for training is mismatched with data used in testing.

In this paper, we analyze ASR performance of the solver systems submitted to the ASpIRE challenge using word error rate (WER) as the performance metric. For a full description of the details of the ASpIRE challenge, see [7]. Toward the goal of evaluating ASR system performance with mismatched training and test conditions, the solver systems were trained on the Fisher conversational telephone training corpus [8]. Solver systems were then evaluated on a new speech corpus, the *Mixer 8 Pilot corpus*, recorded for IARPA by the Linguistic Data Consortium (LDC). The *Mixer 8 Pilot corpus* consists of conversational American English speech recorded via multiple simultaneous microphones spread across seven different rooms in an office-suite environment. Each room exhibited different shapes, sizes, surface properties, and noise sources. The goal of collecting data in this environment was to capture variability that can be observed in real-world speech, and to provide a significant mismatch to the training dataset.

The purpose of our analysis is to identify the factors that contribute to the performance of the solver systems. Namely, we analyze performance as a function of recording room, talker placement, and microphone type and placement. The impact of speech activity detection (SAD) on performance is also investigated.

The rest of this paper is organized as follows: Section 2 gives a brief description of the *Mixer 8 Pilot corpus* used for evaluation and presents the overall performance of the solver systems on the corpus. Section 3 presents the effects of recording conditions on ASR performance. Section 4 evaluates the relationship between SAD and ASR performance. Discussion and conclusions are presented in Section 5.

2. Methods

2.1. Mixer 8 Evaluation corpus

Data evaluated in this paper consists of 120 sessions broken into two different evaluation sets: ASpIRE_single_eval and ASpIRE_multi_eval. Each evaluation set consists of roughly 10 hours of audio, with ASpIRE_single_eval containing one microphone recording per session and ASpIRE_multi_eval containing a selection of six of the eight microphone recordings per session. A simultaneous close-talking telephony channel was recorded as well, but not provided to solvers. The evaluation data was hand-transcribed by Appen Butler Hill for use in scoring.

In addition to the transcripts provided by Appen, LDC provided detailed floor plans and measurements that proved useful in our analysis. Some of these measurements include:

- Distances between talkers and microphones
- Talker positions and angles, relative to the floor plans
- Microphone positions and angles, relative to the floor plans

Special care was taken during the recordings to ensure proper microphone calibration. The microphone gains were calibrated relative to a reference microphone in a special enclosure. This allows a given power measurement relative to full scale to be approximately mapped to dB SPL. Audio calibration sequences (including clicks, tone sweeps, and other stimuli) were recorded in each room on each day and provided to us, along with the sound-meter level recordings of the audio calibration sequences. None of the transcriptions, measurements, or calibration information were provided to the solvers.

2.2. Solver Systems and Performance

Conversational Time Marked files (CTMs) for twelve single-microphone systems and one multi-microphone system were submitted to the ASpIRE challenge. In this paper, we confine our analysis to the single-microphone systems. The systems came from five solvers whose overall system performance is

***Distribution A: Public Release. This work is sponsored by Intelligence Advanced Research Projects Activity (IARPA) under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.**

anonymously summarized in Table 1 with the *n*th scoring solver's *m*th best system receiving the id *n-m*. Primary systems were identified by the solvers as the systems they felt would perform best on the evaluation data.

Solver-System	Primary	WER
1-1	No	43.9
1-2	No	44.0
1-3	Yes	44.3
2-1	Yes	44.3
3-1	Yes	44.8
4-1	No	50.7
4-2	Yes	52.7
4-3	No	52.8
5-1	Yes	53.4
5-2	No	54.1
4-4	No	54.4
4-5	No	54.7

Table 1. Solver-system coding with WER.

3. Effect of Recording Conditions

3.1. Significance of Experimental Setup

The ASpIRE experimental setup varied factors including microphone, speaker position, room, and system, and we wished to investigate whether any of these factors had a significant effect. To explore this question, we ran a multifactor repeated measures Analysis of Variance (ANOVA). The within subject variable was system and the between-subject variables were channel, room, and speaker position. The dependent variable was WER. Table 2 summarizes the ANOVA output, where interactions are specified with colons.

Our ANOVA results indicate that the main effects of room and channel on WER are significant; the interaction effect of room and channel is also significant, suggesting that microphone position as well as microphone audio characteristics affect WER. The main effect of speaker position is not significant, which is as expected since speaker position labels are arbitrary. However, speaker position does have a significant three-way interaction effect with room and channel, again suggesting a relationship between distance between speaker and microphone and WER. The strength of the interaction effects implies that varying room, channel, and speaker position in the experimental setup did have an effect on WER; this will be explored in more detail in the next section.

The system effect of the ANOVA implies that solver systems did differ significantly in WER. Also, the interactions between room and system and channel and system suggest that solver systems had significantly different per-room and per-channel outputs.

Audio file Effect			
Factor	**Degrees of Freedom**	**F Value**	**Pr(>F)**
room	6	9.667	0.0000028***
channel	7	4.395	0.001372**
spkr_pos	2	0.1	0.904875
room:channel	38	2.6	0.002632**
room:spkr_pos	9	0.712	0.694311
channel:spkr_pos	13	1.409	0.204064
room:channel:spkr_pos	9	4.546	0.000519***
System Effect			
Factor	**Degrees of Freedom**	**F Value**	**Pr(>F)**
system	11	130.046	<2.00E-16***
room:system	66	4.638	<2.00E-16***
channel:system	77	1.702	0.000629***
spkr_pos:system	22	0.424	0.990572
room:channel:system	418	0.831	0.968421
room:spkr_pos:system	99	0.812	0.893565
channel:spkr_pos:system	143	0.919	0.720409
room:channel:spkr_pos:system	99	0.684	0.988062

Table 2. Output of repeated measures ANOVA.

3.2. Room Setup Attenuation Metrics

In an effort to understand how to characterize the interaction of room, channel, and speaker position, we investigated the correlation between system WER and various metrics aimed at capturing the effect of microphone and speaker orientation on attenuation of the direct sound. In particular, we examined the *distance attenuation* (due to the distance between the talker and the microphone), the *head directional attenuation* (due to the way the speaker was facing relative to the microphone), and the *microphone directional attenuation* (due to the way the microphone was oriented relative to the speaker).

As the angle of the microphone moves behind the talker, an effect called *head shadow* begins to occur, causing a frequency-dependent attenuation of the signal. Loosely based on a summary of measurements at 2 kHz in [9], we modeled the head directional attenuation in dB as a simple linear function of angle from 0 dB to 10 dB, moving from in-front-of to directly behind the head on the horizontal plane (Figure 1).

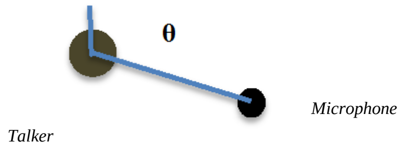

Figure 1. Head directional attenuation: 10*θ/180.

To calculate the microphone directional attenuation, we used linear piecewise functions to approximate the microphone polar attenuation patterns at 1000 Hz. Figure 2 shows the polar attenuation of microphone 5, a Shure MX158, and Figure 3 shows our linear approximation. For omnidirectional microphones, the microphone directional attenuation was zero. The complete set of ASpIRE microphones is listed in Table 3.

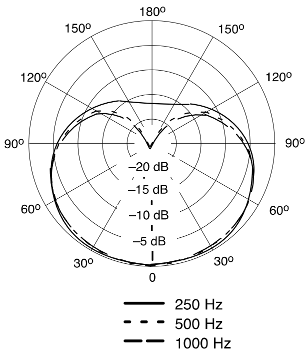

Figure 2. Polar attenuation of microphone 5, a Shure MX158 [10].

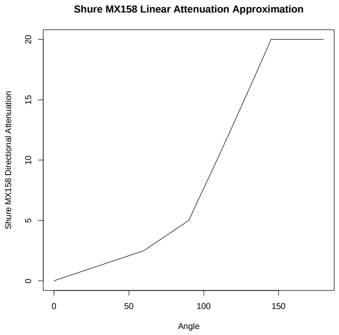

Figure 3. Piecewise linear approximation of microphone 5 attenuation.

Microphone ID	Model	Notes
1	Earthworks M23	Omnidirectional
2	DPA 4090	Omnidirectional
3	Samson SAC02	Directional (Pencil Mic)
4	R0DE NT6	Directional (Miniature)
5	Shure MX185	Directional (Diaphragm condenser)
6	Sony ECMAW3	Omnidirectional (Bluetooth)
7	Canon WM-V1	Omnidirectional (Bluetooth)
8	Audio Technica AT8035	Directional (Shotgun Mic)

Table 3. ASpIRE microphones.

Our final room setup attenuation metric was *total attenuation*, which is the sum of the distance, head directional, and microphone directional attenuation. To evaluate the relationship between the attenuation metrics and WER, we calculated the Spearman's rank correlation coefficient, which was chosen for its ability to detect correlations in nonlinear relationships. The rho values are included in Table 4; with the exception of microphone directional attenuation, all metrics achieve significance at the p ≤ 0.05 level. The positive Spearman's rho values imply that recordings from microphones close to the subject (and oriented toward the subject's mouth) tend to perform better than recordings from microphones further away from the subject.

Total attenuation or the sum of distance attenuation and head directional attenuation show the strongest correlation for all systems, suggesting that taking into account head orientation as well as the distance between speaker and microphone provides an improved model of the effect of distance attenuation on WER. The significance of microphone directional attenuation is less clear, which is notable considering the strong directionality of some of the microphones in the experimental setup. Variation in microphone attenuation at different frequencies could be causing the speech to be filtered rather than broadly attenuated. The significance of microphone orientation for directional microphones on WER is worth further study for future data collections; if orientation is not significant it may not be worth carefully varying, and if it is more significant than it appears under our current investigation, it may be worth recording in more detail.

System	**WER**	**ρ(D. Atten)**	**ρ(HD Atten.)**	**ρ(D + HD Atten)**	**ρ(MD Atten.)**	**ρ(Total Atten.)**
1-1	43.9	**0.309**	**0.196**	**0.349**	0.095	**0.361**
1-2	44.0	**0.342**	**0.191**	**0.375**	0.105	**0.385**
1-3	44.3	**0.36**	**0.202**	**0.395**	0.125	**0.411**
2-1	44.3	**0.366**	**0.218**	**0.431**	0.079	**0.408**
3-1	44.8	**0.286**	**0.232**	**0.355**	0.150	**0.408**
4-1	50.7	**0.305**	**0.227**	**0.376**	0.071	**0.360**
4-2	52.7	**0.273**	**0.259**	**0.37**	0.062	**0.343**
4-3	52.8	**0.277**	**0.257**	**0.373**	0.065	**0.347**
5-1	53.4	**0.388**	**0.194**	**0.419**	0.126	**0.422**
5-2	54.1	**0.38**	**0.185**	**0.409**	0.113	**0.409**
4-4	54.4	**0.282**	**0.244**	**0.364**	0.054	**0.334**
4-5	54.7	**0.27**	**0.253**	**0.357**	0.060	**0.335**

Table 4. Spearman's Rho values between WER and direct signal attenuation metrics (D = distance, HD = head directional, MD = microphone directional). Values that pass a significance test of p ≤ 0.05 are in bold.

3.3. Noise Effects

While our signal attenuation metrics are correlated with solver WER, they do not account for all of the variability. Room noise is known to have an effect on ASR performance and we analyzed the effect of noise on WER across systems. To measure room noise, we calculated the average noise background level and the ratio of speech plus noise power to noise power, which we will hereafter refer to as SNRp. To detect speech and noise, audio files were filtered using A-weighting in Matlab and the noise-only regions were split out using ground-truth SAD labels (see description in section 4).

The power of the resulting noise and non-noise signals were found by calculating the mean of the squared values.

The Spearman's rank correlation coefficients between WER and noise background level, and WER and SNRp, are included in Table 5. Table 5 also includes a column for SAD proportion correct which will be discussed in Section 4. At p ≤ 0.05, background level fails to achieve significance; however, SNRp does achieve significance.

SNRp is strongly negatively correlated to system WER, implying that system performance was affected by ambient. The signal component measured in SNRp contains the direct speech plus all other reflected speech, including energy in the reverberant field, and the noise field. High values of SNRp do not necessarily imply highly intelligible speech, so it is interesting that it is strongly correlated with WER.

SNRp and total attenuation are also significantly correlated as shown in Figure 4; their Spearman rank correlation coefficient is -0.524, p ≤ 0.05. This could imply that manipulating the direct sound in the experimental setup also affected SNRp. However, when we compared the correlation coefficients of a linear model predicting WER from SNRp with the multiple correlation coefficient of a linear model predicting WER from both SNRp and total attenuation (Table 6), we saw only a small difference. This pairs with the stronger correlation of SNRp to WER than total attenuation to WER to suggest that SNRp is related to ASR performance in a manner beyond what is predictable by total attenuation.

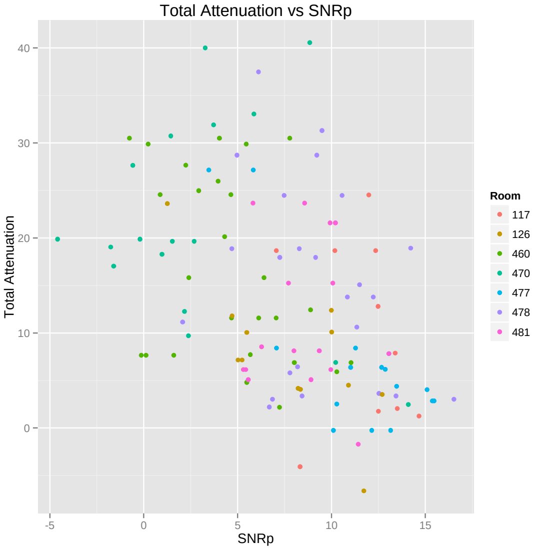

Figure 4. Total Attenuation vs SNRp

System	**WER**	**ρ(Noise Background Level)**	**ρ(SNRp)**	**ρ(Prop. Correct)**
1-1	43.9	0.125	**-0.621**	**-0.550**
1-2	44.0	0.136	**-0.607**	**-0.602**
1-3	44.3	0.151	**-0.622**	**-0.580**
2-1	44.3	0.125	**-0.711**	**-0.424**
3-1	44.8	0.184	**-0.621**	**-0.556**
4-1	50.7	0.115	**-0.587**	**-0.542**
4-2	52.7	0.108	**-0.579**	**-0.531**
4-3	52.8	0.112	**-0.585**	**-0.536**
5-1	53.4	0.106	**-0.577**	**-0.554**
5-2	54.1	0.089	**-0.571**	**-0.581**
4-4	54.4	0.076	**-0.581**	**-0.591**
4-5	54.7	0.076	**-0.577**	**-0.586**

Table 5. Spearman's Rho values between WER and noise background level, SNRp, and proportion correct on SAD judgments. Values that pass a significance test of p ≤ 0.05 are in bold.

System	**WER**	**R(Total Atten.)**	**R(SNRp)**	**R(Total Atten. * SNRp)**
1-1	43.9	**0.372**	**0.672**	**0.691**
1-2	44.0	**0.392**	**0.664**	**0.685**
1-3	44.3	**0.412**	**0.673**	**0.695**
2-1	44.3	**0.399**	**0.724**	**0.736**
3-1	44.8	**0.385**	**0.675**	**0.688**
4-1	50.7	**0.363**	**0.661**	**0.673**
4-2	52.7	**0.354**	**0.653**	**0.665**
4-3	52.8	**0.352**	**0.658**	**0.678**
5-1	53.4	**0.422**	**0.639**	**0.654**
5-2	54.1	**0.401**	**0.631**	**0.640**
4-4	54.4	**0.346**	**0.657**	**0.665**
4-5	54.7	**0.346**	**0.654**	**0.662**

Table 6. Correlation coefficients and coefficients of multiple correlation between WER, total attenuation, and SNRp. All values are significant at p ≤ 0.05.

4. Efficacy of Speech Activity Detection

A key component of the ASpIRE Challenge is the ability of systems to implicitly or explicitly extract speech regions from which to hypothesize transcripts. In this section, we estimate speech activity detection performance (SAD) for each system across a range of speech conditions and relate that detection performance to system WER.

To compute reference SAD values, we divided the audio file into 1 millisecond chunks which we annotated to show whether the chunk occurred during transcribed speech. We computed SAD values for solver system outputs using the same method, and then computed the number of SAD true hits, misses, false alarms, and correct rejections for each system, which are included in Table 7. Three files were omitted from the SAD analysis due to incomplete system outputs from some performers.

Table 5 contains the Spearman's rank correlation coefficient for the per-file relationship between system WER and the proportion of correct SAD judgments (true hits + correct rejections). At a p-value of 0.05, system WER and correct SAD judgments are significantly correlated for all single-microphone systems.

System	WER	Average Prop. Correct	Average Prop. Miss	Average Prop. False Alarm
1-1	43.9	0.848	0.131	0.022
1-2	44.0	0.828	0.154	0.018
1-3	44.3	0.829	0.153	0.018
2-1	44.3	0.848	0.133	0.019
3-1	44.8	0.841	0.149	0.010
4-1	50.7	0.785	0.210	0.005
4-2	52.7	0.795	0.195	0.010
4-3	52.8	0.796	0.195	0.010
5-1	53.4	0.789	0.191	0.020
5-2	54.1	0.786	0.192	0.022
4-4	54.4	0.782	0.209	0.009
4-5	54.7	0.78	0.211	0.009

Table 7. System SAD performance

Focusing on the proportion of correct SAD judgments (which we will refer to as the “SAD accuracy” below), we include a plot of average system SAD accuracy against WER in Figure 5. The systems show two clusters, with top performing systems also achieving the best SAD performance. Although we cannot claim that this shows that SAD performance determines WER, it implies that solvers might try improving their SAD systems as they seek to improve their WER. Figure 6 shows that although all submitted systems had very low false alarm SAD rates, top-performing systems cluster in their SAD operating point at a lower miss rate, suggesting a possible direction for some solvers to explore in improving their performance on the ASpIRE task.

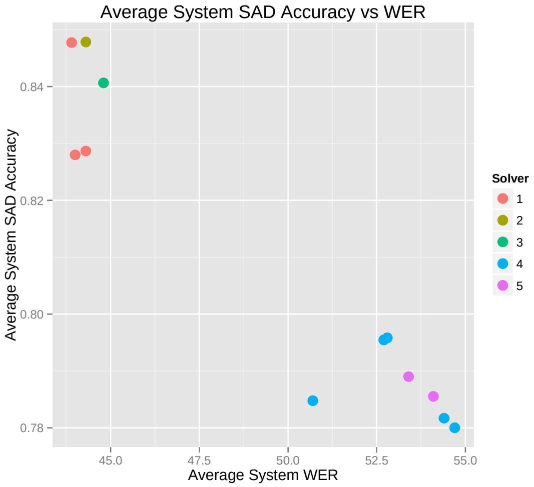

Figure 5. Average system SAD accuracy versus system WER

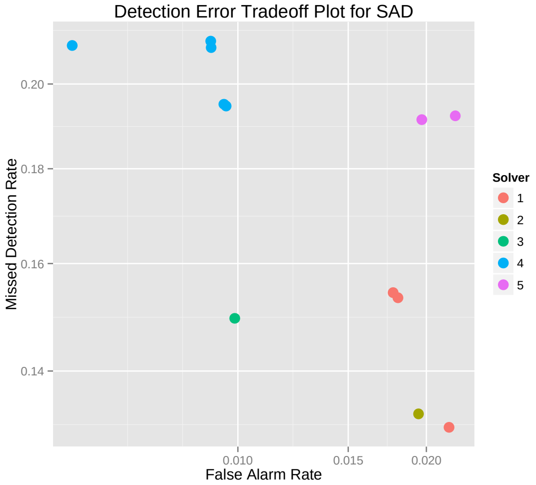

Figure 6. DET plot for system average SAD performance

In an effort to investigate whether high SNR conditions might be correlated to degradations in SAD performance as well as WER, we computed the Spearman’s rank correlation coefficients between SNRp and solver SAD. Table 8 shows that the two are indeed significantly correlated. These results further suggest that if solvers improve their SAD systems on higher SNR conditions, they might expect to see gains in WER.

System	WER	ρ(SNRp, SAD Performance)
1-1	43.9	**0.463**
1-2	44.0	**0.473**
1-3	44.3	**0.457**
2-1	44.3	**0.397**
3-1	44.8	**0.456**
4-1	50.7	**0.469**
4-2	52.7	**0.464**
4-3	52.8	**0.464**
5-1	53.4	**0.522**
5-2	54.1	**0.531**
4-4	54.4	**0.470**
4-5	54.7	**0.464**

Table 8. Spearman’s rho for the relationship between SNRp and system SAD performance. All values are significant at $p \leq 0.05$.

5. Discussion and Conclusion

From the analysis conducted in this paper, we observed that the experimental factors of room, channel, and speaker position varied in the ASpIRE challenge have a significant interactional effect on ASR performance.

We demonstrated that the total distance between microphone and speaker, taking into account orientation of both, is better-correlated with ASR system performance than simpler distance metrics. We showed that room noise measured through SNRp has a strong correlation to degradations in WER, and that total attenuation is strongly correlated with SNRp. Finally, we showed that system SAD performance also shows a strong correlation to WER as well as to SNRp and naturally partitions the better-performing systems in the ASpIRE challenge from the rest. This result implies that solvers might perform better in conditions similar to those in the ASpIRE challenge by doing further work to improve their SAD systems under high SNR.

Work remains to be done to further investigate the effect of microphone direct signal attenuation on ASR performance in order to inform the care with which it is treated in future data collections. Additionally, although top solver systems were very similar in WER, the solvers used highly varied ASR algorithms. It would be interesting to do a more detailed analysis of differences (if any) in how various ASR techniques are affected by the experimental conditions varied in the ASpIRE challenge. Finally, further analysis could be done to see what gain in WER would be possible from a combined system.

6. Acknowledgments

The authors would like to thank Mary Harper at IARPA for her creation and overall management of the ASpIRE challenge, and her many useful comments on drafts of this paper. Also, Michael Brandstein, Joe Campbell, and Jon Fiscus for their useful advice.

7. References

[1] Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., "The Reverb Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," in Proceedings of WASPAA, 2013.

[2] Wolfel, M. and McDonough, J., “Distant Speech Recognition,” Wiley, 2009.

[3] Hain, T., Wan, V., Burget, L., Karafiat, M., Dines, J., Vepa, J., Garau, G., Lincoln, M., "The AMI System for the Transcription of Speech in Meetings," in Proceedings of ICASSP, 2007.

[4] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C., “The ICSI Meeting Corpus”, in Proceedings of ICASSP, 2003.

[5] Parthasarathi, S. H. K., Chang, S. Y., Cohen, J., Morgan, N., and Wegmann, S., “The Blame Game in Meeting Room ASR: An Analysis of Feature Versus Model Errors in Noisy and Mismatched Conditions”, in Proceedings of ICASSP, 2013.

[6] Lincoln, M., McCowan, I., Vepa, J. Maganti, H. K., “The Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments”, IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.

[7] Harper, M., The Automatic Speech Recognition in Reverberant Environments (ASpIRE) Challenge. 2015. Submitted to ASRU.

[8] Cieri, C., Miller, D., Walker, K., “The Fisher Corpus, A Resource for the Next Generations of Speech-to-Text”, in Proceedings 4th International Conference on Language Resources and Evaluation, 2004.

[9] Kompis, M. and Dillier, N., “Simulating Transfer Functions in a Reverberant Room Including Source Directivity and Head-Shadow Effects”, Journal Acoustical Society of America 93 (5), 2779-2787, 1993.

[10] Shure Model MX100 Microphones Specification Sheet. Retrieved from http://cdn.shure.com/specification_sheet/upload/42/us_pro_mx183_184_185_specsheet.pdf on June 17, 2015.